Ecological Informatics 23 (2014) 69–82

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

A model for environmental data extraction from multimedia and its evaluation against various chemical weather forecasting datasets

Anastasia Moumtzidou [a,*], Victor Epitropou [b], Stefanos Vrochidis [a], Kostas Karatzas [b], Sascha Voth [c], Anastasios Bassoukos [b], Jürgen Moßgraber [c], Ari Karppinen [d], Jaakko Kukkonen [d], Ioannis Kompatsiaris [a]

[a] Information Technologies Institute, Centre for Research and Technology Hellas, Greece
[b] Informatics Systems and Applications Group, Aristotle University of Thessaloniki, Greece
[c] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Germany
[d] Finnish Meteorological Institute, Helsinki, Finland

ARTICLE INFO

Article history:
Received 31 January 2013
Received in revised form 12 July 2013
Accepted 20 August 2013
Available online 4 September 2013

Keywords:
Air quality
Heatmap
Image processing
OCR
Environmental
Multimedia

ABSTRACT

Environmental data analysis and information provision are considered of great importance for people, since environmental conditions are strongly related to health issues and directly affect a variety of everyday activities. Nowadays, there are several free web-based services that provide environmental information in several formats with map images being the most commonly used to present air quality and pollen forecasts. This format, despite being intuitive for humans, complicates the extraction and processing of the underlying data. Typical examples of this case are the chemical weather forecasts, which are usually encoded heatmaps (i.e. graphical representation of matrix data with colors), while the forecasted numerical pollutant concentrations are commonly unavailable. This work presents a model for the semi-automatic extraction of such information based on a template configuration tool, on methodologies for data reconstruction from images, as well as on text processing and Optical Character Recognition (OCR). The aforementioned modules are integrated in a standalone framework, which is extensively evaluated by comparing data extracted from a variety of chemical weather heat maps against the real numerical values produced by chemical weather forecasting models. The results demonstrate a satisfactory performance in terms of data recovery and positional accuracy.

1. Introduction

The analysis of environmental data and the generation, combination and reuse of related information, such as air pollutant concentrations, is of particular interest for people. Environmental status information (in particular, the concentration of certain pollutants in the air) is considered to be correlated with a series of health issues, such as cardiovascular and respiratory diseases, it directly affects several outdoor activities (e.g. commuting, sports, trip planning, agriculture) and therefore it is strongly related to the overall quality of life. In addition, the analysis of environmental information is often a prerequisite for the fulfillment of legal mandates on the management and preservation of environmental quality, according to the EU's and other legal frameworks (Karatzas and Moussiopoulos, 2000). With a view to offering personalized decision support services for people based on environmental information regarding their everyday activities (Wanner et al., 2012) and supporting environmental experts in air quality preservation tasks, there is a need to extract, combine and compare complementary and competing environmental information from several resources in order to generate more reliable and cross-validated information on the environmental conditions. One of the main steps towards this goal is the environmental information extraction from heterogeneous resources.

Environmental observations are automatically performed by specialized instruments, hosted in stations established by environmental organizations, while the forecasts, which are used to foretell weather conditions, the levels of pollution or pollen concentration in areas of interest, are provided by environmental prediction models, the output of which are gridded numerical data, henceforth referred to as 'actual' or 'original' data. In practice only a few of the data providers make available to the public some means of access to their actual (numerical) forecast data, while the majority publishes the results in the form of preprocessed images, that address specific environmental pressures (like air pollution concentrations), for specific temporal scales (usually in the order of hours or days), and for specific geographical areas of interest. However, even if the original data values of environmental information had been available, these would commonly be presented in various technical formats, using various coordinates and spatial resolutions, different units, and several other choices (e.g., Kukkonen et al.,

* Corresponding author at: Centre for Research and Technology Hellas, Information Technologies Institute, 6th km Charilaou-Thermi Road, P.O. Box 60361, 57001 Thermi, Thessaloniki, Greece. Tel.: +30 2311257746.
E-mail addresses: moumtzid@iti.gr (A. Moumtzidou), vepitrop@isag.meng.auth.gr (V. Epitropou), stefanos@iti.gr (S. Vrochidis), kkara@eng.auth.gr (K. Karatzas), sascha.voth@iosb.fraunhofer.de (S. Voth), abas@isag.meng.auth.gr (A. Bassoukos), juergen.mossgraber@iosb.fraunhofer.de (J. Moßgraber), ari.karppinen@fmi.fi (A. Karppinen), jaakko.kukkonen@fmi.fi (J. Kukkonen), ikom@iti.gr (I. Kompatsiaris).

1574-9541/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecoinf.2013.08.003

2012). It can therefore be a laborious task to convert these data files to the same harmonized format, for inter-comparison purposes. Consequently, the main sources of environmental information for everyday use are web portals and sites, which provide a variety of information of diverse spatial and temporal nature. Although the weather forecasts are usually presented in textual format (Moumtzidou et al., 2012b), important environmental information such as the air quality and pollen forecasts is encoded in multimedia formats (Karatzas, 2005). Specifically, the vast majority of such environmental data are published as static heatmaps (i.e. graphical representation of matrix data with colors), or as sequences of heatmaps (time-lapse animations). A characteristic example of a heatmap is presented in Fig. 1 (generated by the SILAM model, courtesy of FMI). However, since this information comes from different providers and is presented in a variety of not intercomparable and compatible visual forms, it is not possible to directly combine them and compile a synthetic service that takes into account all available data sources. In order to deal with this problem, it is necessary to design and develop a model that is capable of extracting environmental information from heatmaps and translate them to a structured numerical format. The processing of images for their conversion into numerical data would comprise the core of environmental data recovery techniques, at least in the air pollution and the pollen concentration domains.

In this context, this paper addresses the extraction of air quality and pollen forecasts from heatmaps, by proposing a semi-automatic framework, which consists of three main components: an annotation tool for administrative user intervention used for generating configuration templates for each heatmap, an Optical Character Recognition (OCR) and text processing module used for fetching text information embedded in the image and making the necessary corrections, as well as the AirMerge heatmap processing module (Epitropou et al., 2011) that allows for the automatic harvesting, annotation, harmonization and reconversion of heatmaps into numerical data. The framework is evaluated against the AirMerge system and various chemical weather forecast datasets. It should be highlighted that the AirMerge system, per-se, does not include an automated annotation process, therefore any heatmap harvesting and parsing procedure must be manually scripted, even though the programmatic generation of certain types of highly repetitive scripts e.g. to handle series of images from one same provider, is possible. On the contrary, the proposed framework aims at automating this scripting process, by generating the configuration scripts required by AirMerge on a per-case basis via optical heatmap analysis, the use of graphical templates and machine automation. The results of the resulting scripts are then compared to those obtained by using the best manually configured AirMerge scripts for a given heatmap template, and the differences in their setup and final data extraction results are discussed.

The contribution of this work is a novel framework that integrates multimedia annotation and processing modules, in order to allow for the semi-automatic extraction of air quality and/or pollen forecast data presented in heatmaps. Specifically, this framework integrates multimedia configuration components (annotation tool), advanced systems for heatmap image processing (AirMerge) and optimized OCR techniques. These modules are integrated in a standalone, user-based interface that allows for template-based customization of heatmaps and thus assists in handling several formats of heatmaps. This paper substantially extends the works presented in Moumtzidou et al. (2012a) and Vrochidis et al. (2012), which have demonstrated the initial results of this framework, by providing an extensive evaluation, which includes a comparative study of the proposed framework against the manually configured AirMerge system and real numerical data provided by forecast models for a variety of providers.

This paper is structured as follows: Section 2 presents the previous research on heatmap analysis and content extraction, Section 3 describes the results of studies on the presentation format of environmental

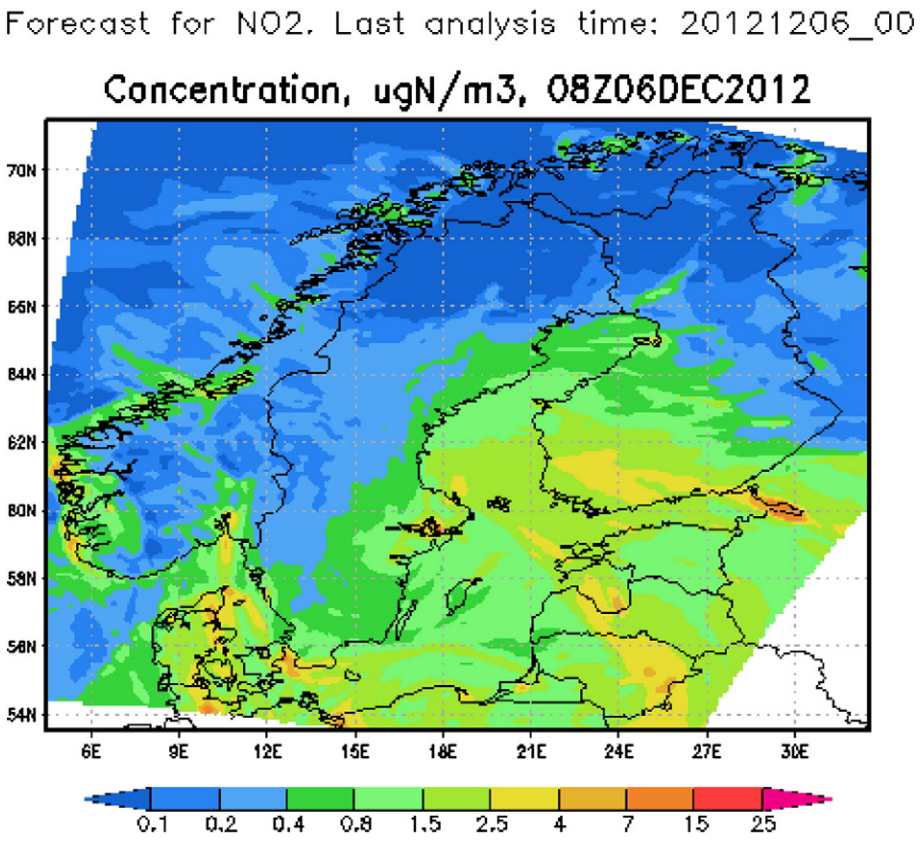

Fig. 1. An example of an air quality heatmap: the forecast of NO_2 concentrations ($\mu g/m^3$) at 8 UTC time of 6 December 2012, using the SILAM chemical transport model.